Lasso Regression

Nipun Batra

IIT Gandhinagar

August 1, 2025

Lasso Regression

- LASSO $\longrightarrow$ Least absolute shrinkage and selection operator

Lasso Regression

- LASSO $\longrightarrow$ Least absolute shrinkage and selection operator
- Popular as it leads to a sparse solution.

Constructing the Objective Function

- Find a θ_{opt} such that

$$\theta_{\text{opt}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) : \ ||\boldsymbol{\theta}||_1 < s \qquad (1)$$

Constructing the Objective Function

- Find a θ_{opt} such that

$$\boldsymbol{\theta}_{\text{opt}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) : \ ||\boldsymbol{\theta}||_1 < s \tag{1}$$

- Using KKT conditions

$$\boldsymbol{\theta}_{\text{opt}} = \underbrace{\arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 ||\boldsymbol{\theta}||_1}_{\text{convex function}} \tag{2}$$

Solving the Objective

- Since $||\boldsymbol{\theta}||_1$ is not differentiable, we cannot solve,

$$\frac{\partial(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})+\delta^2||\boldsymbol{\theta}||_1}{\partial\boldsymbol{\theta}}=0 \tag{3}$$

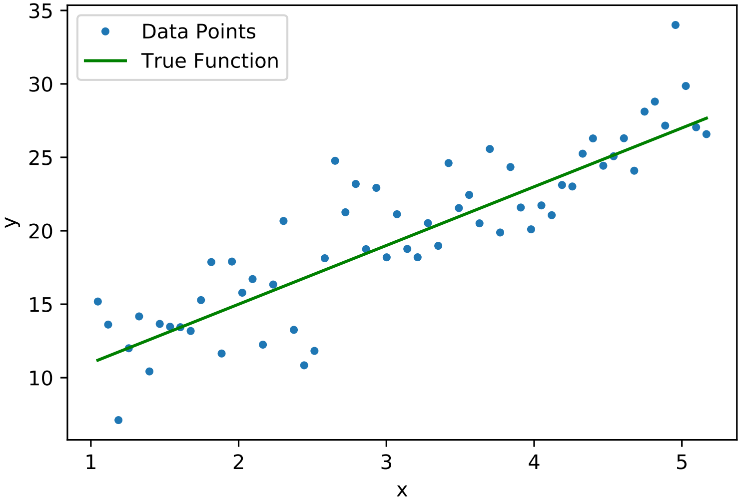
Solving the Objective

- Since $||\boldsymbol{\theta}||_1$ is not differentiable, we cannot solve,

$$\frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2||\boldsymbol{\theta}||_1}{\partial\boldsymbol{\theta}} = 0 \tag{3}$$

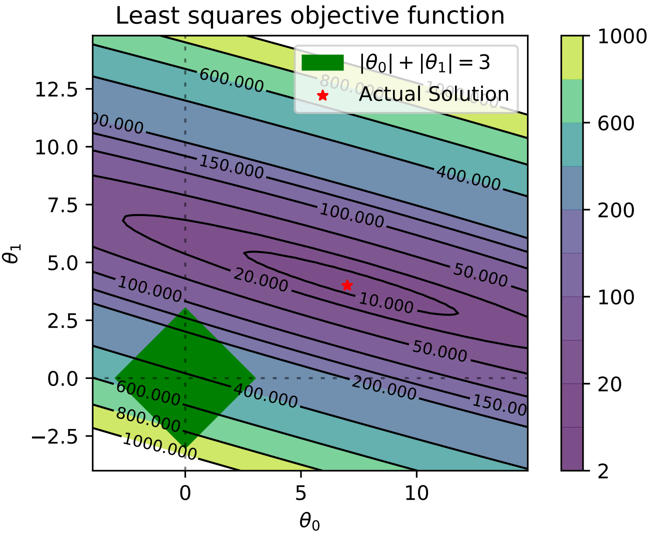
- How to Solve? Use coordinate descent!

Sample Dataset



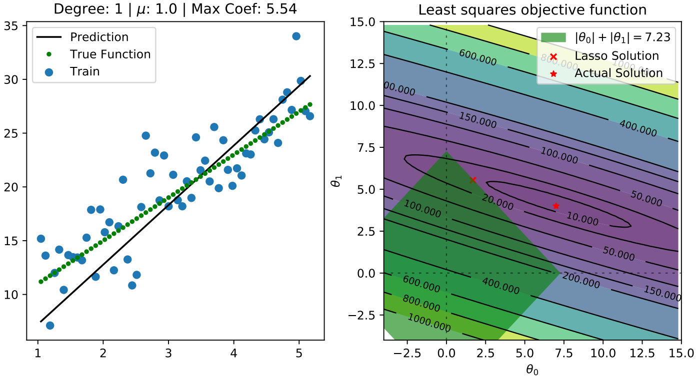
$y = 4x + 7$

Geometric Interpretation



Lasso regression

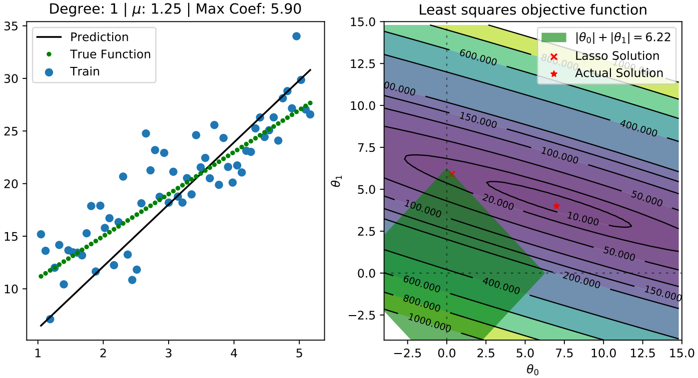
Effect of μ - Regularization of Parameters



$$\mu = 1.0$$

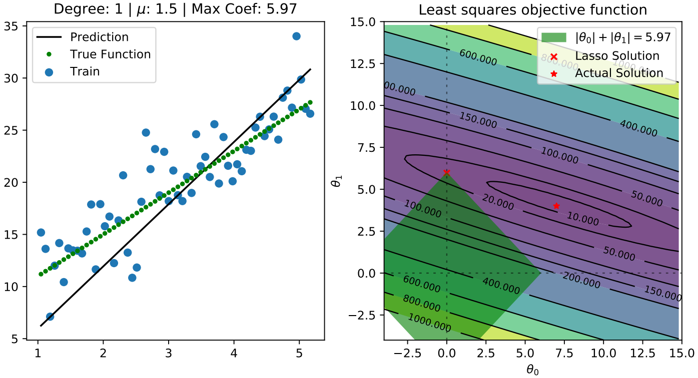
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters



$\mu = 1.25$
(on the *Sample Dataset*)

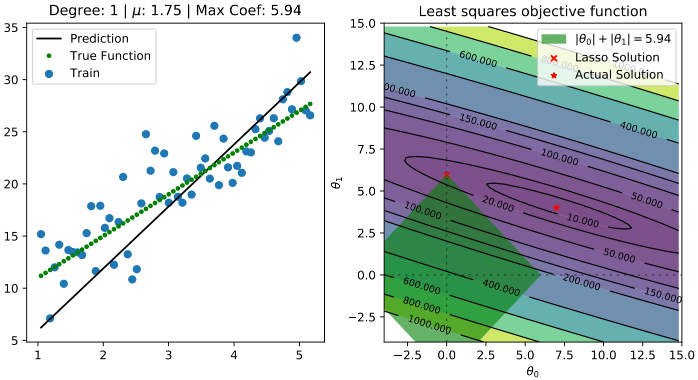
Effect of μ - Regularization of Parameters



$$\mu = 1.5$$

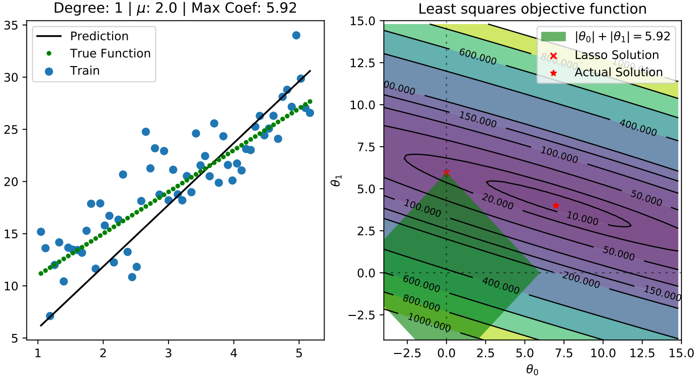
(on the *Sample Dataset*)

Effect of μ - Regularization of Parameters



$$\mu = 1.75$$
(on the *Sample Dataset*)

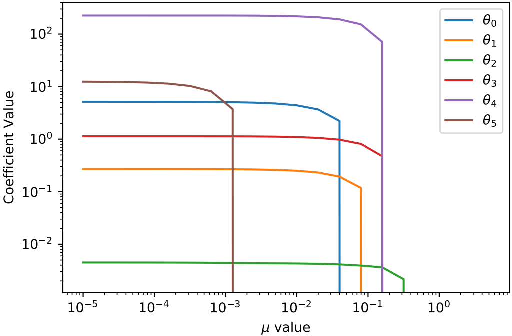
Effect of μ - Regularization of Parameters



$$\mu = 2.0$$

(on the *Sample Dataset*)

Regularization path of lasso regression



Regularization path of θ_i

LASSO and feature selection

- LASSO inherently does feature selection!

LASSO and feature selection

- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.

LASSO and feature selection

- LASSO inherently does feature selection!
- Sets coefficients of “less important” features to zero.
- Sparse and memory efficient and often more interpretable models.

Subgradient

- Generalises gradient to convex but non-differentiable problems

Subgradient

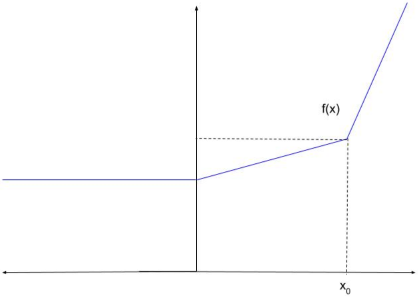
- Generalises gradient to convex but non-differentiable problems
- Examples:

Subgradient

- Generalises gradient to convex but non-differentiable problems
- Examples:
  - $f(x) = |x|$

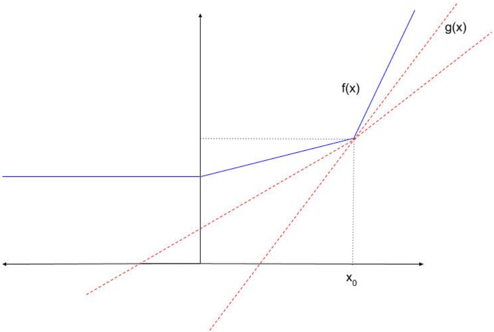
Task at hand

- TASK: find derivative of $f(x)$ at $x = x_0$



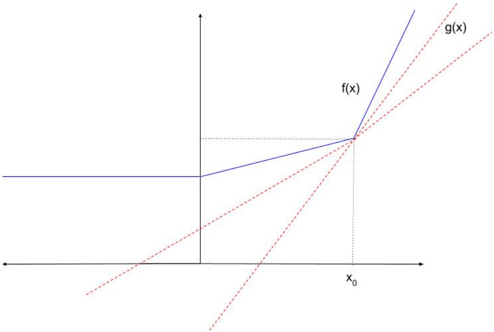
Solution

- Construct a differentiable $g(x)$



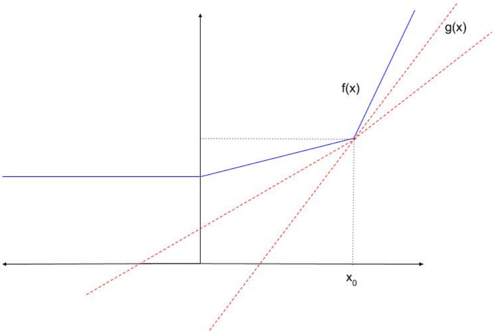
Solution

- Construct a differentiable $g(x)$
  - Intersecting $f(x)$ at $x = x_0$



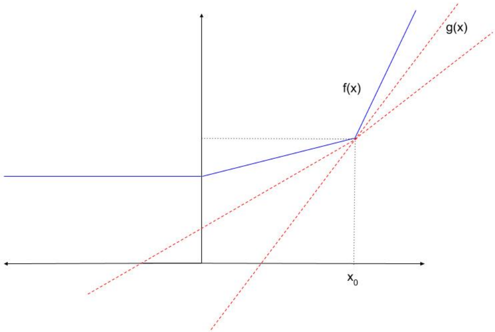
Solution

- Construct a differentiable $g(x)$
    - Intersecting $f(x)$ at $x = x_0$
    - Below or on $f(x)$ for all x



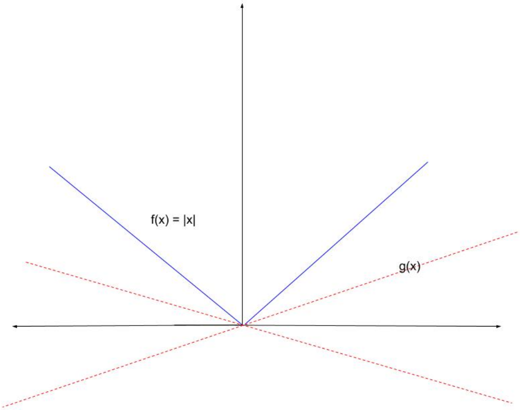
Solution

- Compute slope of $g(x)$ at $x = x_0$



Another Example: $f(x) = |x|$

- Subgradient of $f(x)$ belongs to $[-1, 1]$



Coordinate Descent

- Another optimisation method (akin to gradient descent)

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$
- Key idea: Sometimes difficult to find minimum for all coordinates

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate

Coordinate Descent

- Another optimisation method (akin to gradient descent)
- Objective: $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$
- Key idea: Sometimes difficult to find minimum for all coordinates
- ..., but, easy for each coordinate
- turns into a one-dimensional optimisation problem

Coordinate Descent

- Picking next coordinate:

Coordinate Descent

- Picking next coordinate:

Coordinate Descent

- Picking next coordinate: random, round-robin
- No step-size to choose!

Coordinate Descent

- Picking next coordinate: random, round-robin
- No step-size to choose!
- Converges for Lasso objective

Coordinate Descent : Example

Learn $y = \theta_0 + \theta_1 x$ on following dataset, using coordinate descent where initially $(\theta_0, \theta_1) = (2, 3)$ for 2 iterations.

x	**y**
1	1
2	2
3	3

Coordinate Descent : Example

Our predictor, $\hat{y} = \theta_0 + \theta_1 x$

Error for i^{th} datapoint, $\epsilon_i = y_i - \hat{y}_i$
$\epsilon_1 = 1 - \theta_0 - \theta_1$
$\epsilon_2 = 2 - \theta_0 - 2\theta_1$
$\epsilon_3 = 3 - \theta_0 - 3\theta_1$

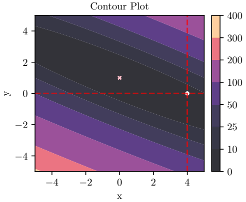
$$\mathrm{MSE} = \frac{\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2}{3} = \frac{14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1}{3}$$

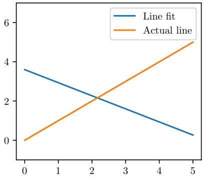
Iteration 0

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$

Iteration 0

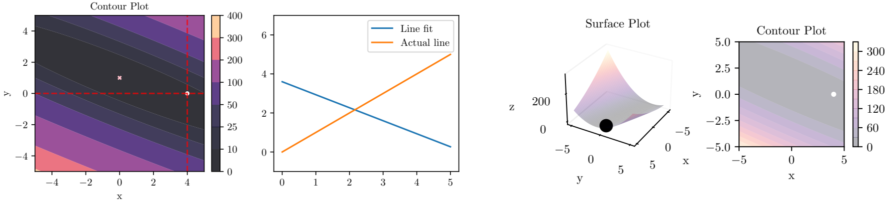
MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$





Iteration 0

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$



Coordinate Descent : Example

Iteration 1

INIT: $\theta_0 = 2$ and $\theta_1 = 3$

$\theta_1 = 3$ optimize for θ_0

Coordinate Descent : Example

Iteration 1

INIT: $\theta_0 = 2$ and $\theta_1 = 3$

$\theta_1 = 3$ optimize for θ_0

$\frac{\partial \operatorname{MSE}}{\partial \theta_0} = 6\theta_0 + 24 = 0$

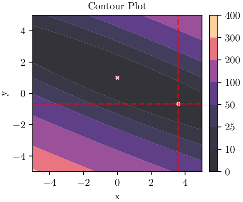
$\theta_0 = -4$

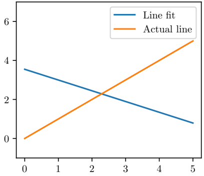
Iteration 1

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$

Iteration 1

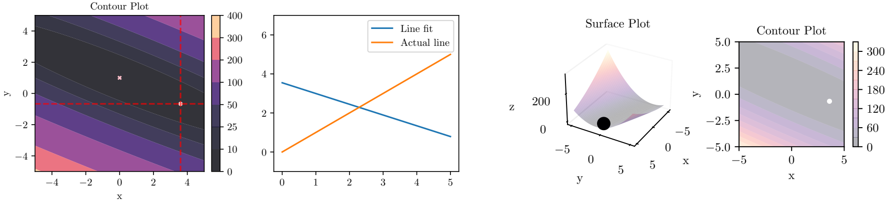
MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$





Iteration 1

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$



Coordinate Descent : Example

Iteration 2

INIT: $\theta_0 = -4$ and $\theta_1 = 3$

$\theta_0 = -4$ optimize for θ_1

Coordinate Descent : Example

Iteration 2

INIT: $\theta_0 = -4$ and $\theta_1 = 3$

$\theta_0 = -4$ optimize for θ_1

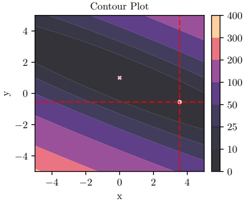
$\theta_1 = 2.7$

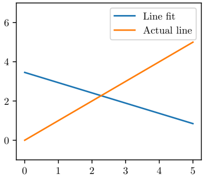
Iteration 2

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$

Iteration 2

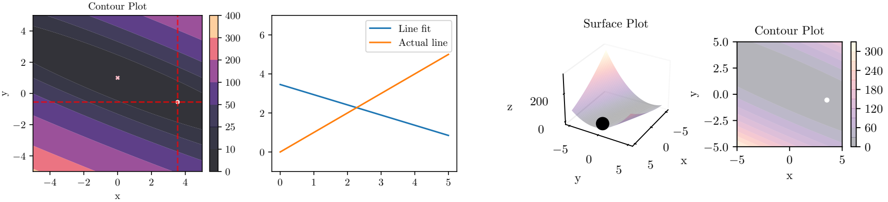
MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$





Iteration 2

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$



Coordinate Descent : Example

Iteration 3

INIT: $\theta_0 = -4$ and $\theta_1 = 2.7$

$\theta_1 = 2.7$ optimize for θ_0

Coordinate Descent : Example

Iteration 3

INIT: $\theta_0 = -4$ and $\theta_1 = 2.7$

$\theta_1 = 2.7$ optimize for θ_0

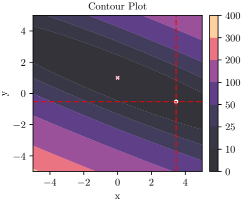
$\theta_0 = -3.4$

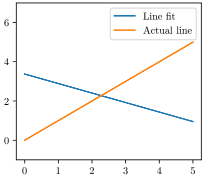
Iteration 3

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$

Iteration 3

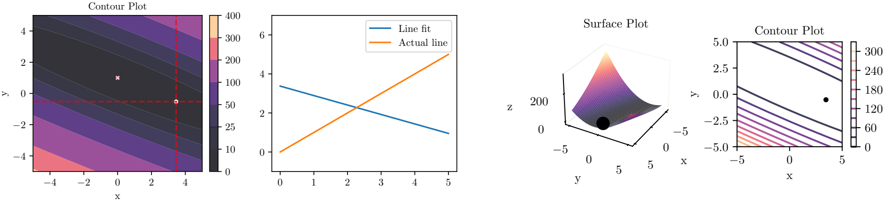
MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$





Iteration 3

MSE = $\frac{1}{3}(14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1)$



Coordinate Descent for Unregularised Regression

- Express error as a difference of y_i and $\hat{y}_i$

$$\hat{y}_i = \sum_{j=0}^{d} \theta_j x_i^j = \theta_0 x_i^0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \ldots + \theta_d x_i^d \quad (4)$$

$$\epsilon_i = y_i - \hat{y}_i = y_i - \theta_0 x_i^0 - \theta_1 x_i^1 - \ldots - \theta_d x_i^d = y_i - \sum_{j=0}^{d} \theta_j x_i^j \quad (5)$$

Coordinate Descent for Unregularised regression

$$\sum_{i=1}^{n} \epsilon^2 = \mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

Coordinate Descent for Unregularised regression

$$\sum_{i=1}^{n} \epsilon^2 = \mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \theta_d x_i^d \right) \right)^2$$

$$\frac{\partial \, \mathrm{RSS}\left(\theta_j\right)}{\partial \theta_j} = 2 \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \ldots \right) \right) \left(-x_i^j \right)$$

Coordinate Descent for Unregularised regression

$$\sum_{i=1}^{n} \epsilon^2 = \mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \theta_d x_i^d\right)\right)^2$$

$$\frac{\partial \, \mathrm{RSS}\left(\theta_j\right)}{\partial \theta_j} = 2 \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \ldots\right)\right) \left(-x_i^j\right)$$

$$= 2 \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_d x_i^d\right)\right) \left(-x_i^j\right) + 2 \sum_{i=1}^{n} \theta_j (x_i^j)^2$$

Coordinate Descent for Unregularised regression

$$\sum_{i=1}^{n} \epsilon^2 = \mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \theta_d x_i^d\right)\right)^2$$

$$\frac{\partial\, \mathrm{RSS}\left(\theta_j\right)}{\partial \theta_j} = 2 \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_j x_i^j + \ldots\right)\right) \left(-x_i^j\right)$$

$$= 2 \sum_{i=1}^{n} \left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_d x_i^d\right)\right) \left(-x_i^j\right) + 2 \sum_{i=1}^{n} \theta_j (x_i^j)^2$$

where:

$$\hat{y}_i^{(-j)} = \theta_0 x_i^0 + \ldots + \theta_d x_i^d$$

is $\hat{y}_i$ without θ_j

Coordinate Descent for Unregularised regression

$$\text{Set } \frac{\partial \operatorname{RSS}\left(\theta_j\right)}{\partial \theta_j} = 0$$

$$\theta_j = \sum_{i=1}^{n} \frac{\left(y_i - \left(\theta_0 x_i^0 + \ldots + \theta_d x_i^d\right)\right)\left(x_i^j\right)}{\left(x_i^j\right)^2} = \frac{\rho_j}{z_j}$$

$$\rho_j = \sum_{i=1}^{n} x_i^j \left(y_i - \hat{y}_i^{(-j)}\right) \quad \text{and} \quad z_j = \sum_{i=1}^{n} \left(x_i^j\right)^2$$

z_j is the squared of ℓ_2 norm of the j^{th} feature

Coordinate Descent for Lasso Regression

$$\text{Minimise} \underbrace{\sum_{i=1}^{n} \epsilon^2 + \delta^2 \left\{|\theta_0| + |\theta_1| + \ldots \left|\theta_j\right| + \ldots |\theta_d|\right\}}_{\text{LASSO OBJECTIVE}}$$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \frac{\partial}{\partial \theta_j} \left|\theta_j\right|$$

$$\frac{\partial}{\partial \theta_j} \left|\theta_j\right| = \begin{cases} 1 & \theta_j > 0 \\ [-1, 1] & \theta_j = 0 \\ -1 & \theta_j < 0 \end{cases}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

Coordinate Descent for Lasso Regression

- **Case 1:** $\theta_j > 0$

$$-2\rho_j + 2\theta_j z_j + \delta^2 = 0$$

$$\theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

$$\rho_j > \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j - \frac{\delta^2}{2}}{z_j}$$

- **Case 2:** $\theta_j < 0$

$$\rho_j < \frac{\delta^2}{2} \Rightarrow \theta_j = \frac{\rho_j + \delta^2/2}{z_j} \tag{6}$$

Coordinate Descent for Lasso Regression

- **Case 3:** $\theta_j = 0$

$$\frac{\partial}{\partial \theta_j}(\text{LASSO OBJECTIVE}) = -2\rho_j + 2\theta_j z_j + \delta^2 \underbrace{\frac{\partial}{\partial \theta_j}|\theta_j|}_{\text{[-1,1]}}$$

$$\in \underbrace{[-2\rho_j - \delta^2, -2\rho_j + \delta^2]}_{\{0\} \text{ lies in this range}}$$

$$-2\rho_j - \delta^2 \leq 0 \text{ and } -2\rho_j + \delta^2 \geq 0$$

$$-\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \Rightarrow \theta_j = 0$$

Summary of Lasso Regression

$$\theta_j = \begin{bmatrix} \frac{\rho_j + \frac{\delta^2}{2}}{z_j} & \text{if} & \rho_j < -\frac{\delta^2}{2} \\ 0 & \text{if} & -\frac{\delta^2}{2} \leq \rho_j \leq \frac{\delta^2}{2} \\ \frac{\rho_j - \frac{\delta^2}{2}}{z_j} & \text{if} & \rho_j > \frac{\delta^2}{2} \end{bmatrix} \tag{7}$$